Differentiating Operator Skill during Routine Fetal Ultrasound Scanning using Probe Motion Tracking

Yipei Wang[1], Richard Droste[1], Jianbo Jiao[1], Harshita Sharma[1], Lior Drukker[2], Aris T. Papageorghiou[2], and J. Alison Noble[1]

[1] Institute of Biomedical Engineering, University of Oxford, Oxford, UK
[2] Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK
yipei.wang@eng.ox.ac.uk

Abstract. In this paper, we consider differentiating operator skill during fetal ultrasound scanning using probe motion tracking. We present a novel convolutional neural network-based deep learning framework to model ultrasound probe motion in order to classify operator skill levels, that is invariant to operators' personal scanning styles. In this study, probe motion data during routine second-trimester fetal ultrasound scanning was acquired by operators of known experience levels (2 newly-qualified operators and 10 expert operators). The results demonstrate that the proposed model can successfully learn underlying probe motion features that distinguish operator skill levels during routine fetal ultrasound with 95% accuracy.

Keywords: Operator skill · Probe motion · Fetal ultrasound.

1 Introduction

Ultrasound is a relatively low-cost medical imaging modality, convenient, non-invasive, yields real-time results and is widely considered to be safe. Therefore, it is the primary modality for pregnancy imaging used to assess the fetal anatomy, development and growth [5]. During an ultrasound scan, the operator is required to skillfully manipulate the ultrasound probe to acquire a series of standard anatomical planes for diagnostic interpretation or biometric measurements. Even minor probe movements could cause distortion of the viewed anatomy or largely influence the localization of the standard plane. Therefore, sonography is well recognized as a task that is difficult to learn. It takes a significant amount of time to train fetal ultrasound sonographers before reaching competency [15].

Despite the importance of ultrasound in obstetrics and difficulty in operator training, there is little quantitative research studying the process of operator training and operator skill assessment for obstetric scanning. Protocols and guidelines [11, 12] exist for scanning but there is no universal objective sonographer skill assessment standard for fetal ultrasound scanning. Operators are

trained and assessed under supervision and given feedback. Such a process is prone to human bias and might not be consistent across different settings [7]. Therefore, it would be practically useful to design new and automated objective approaches to support new trainees and inform them once they reach expert competency.

One approach recently introduced in the related area of surgical skill assessment is objective computer-aided technical skill evaluation [14]. Researchers have developed machine learning methods that use tool motion [2, 9, 18], surgeon/tool video [17, 19, 20], surgeons' eye-gaze [3], and the combination of these data to model a surgical procedure. In this study, we look at sonography from a similar perspective of surgical data science, where machine learning methods are used to classify skills and support training. We also factor in that ultrasound scanning is highly operator-dependent [4]. The underlying features of operator skill and personal style are entangled, which makes the model training process rather difficult. Therefore, recognising operator style-agnostic skills is crucial to building models that characterise operator skill.

Our contributions are as follows: (1) we propose a deep learning framework to differentiate operator skill based on probe motion data; (2) we constrain the training of networks to make the learnt models invariant to operators' personal scanning styles. Experiments show that the proposed framework is capable of learning operator-invariant features that distinguish different skill levels.

2 Method

2.1 Data Acquisition and Annotation

Probe motion data was acquired as part of the PULSE [13] study. This study was approved by the UK Research Ethics Committee (Reference 18/WS/0051), and written informed consent was given by all participating pregnant women. Sonographers also consented to participate in the study at the outset, but did not have any visual or other signal to know that the probe motion tracking device was functioning. The ultrasound scans were performed by qualified sonographers and fetal medicine doctors (collectively referred to as operators in this paper) using a General Electric Voluson E8 ultrasound machine. The ultrasound machine was equipped with a standard curvilinear (C2-9-D, C1-5-D) and 3D/4D (RAB6-D) probes. The selection of suitable motion tracking methods was limited. We could not use any electromagnetic trackers or optical trackers due to safety reasons in the scan room and not wide enough field of view, respectively. Hand motion video recording was also not possible because of privacy of patients. In this study, an inertial measurement unit was rigidly attached to the probe to record the probe motion data.

The research goal of this work was to understand what can be learned from motion alone independent of video. Ultrasound video was only used to identify the motion data segments corresponding to episodes of fetal brain scanning. We selected the collection of motion segments with the following criteria: 1) the video

and motion data is well synchronized; 2) the scanner is not in image freeze mode during the segment; 3) the segment has been assigned an anatomy label (the brain). Scanning parameters were automatically extracted for each video frame in the full-length scan video using optical character recognition. This selection resulted in 396 motion segments of fetal brain scanning from 229 full-length scans performed by 12 operators. The original raw motion data was sampled at 400Hz. To reduce the data dimension while preserving useful information, the data was downsampled to 30Hz to match with the ultrasound video sampling rate.

We define operators with more than two years of experience of fetal anomaly screening ultrasound scanning as "expert" level, and those with less than two years of experience as "newly qualified". Data for 2 newly qualified and 10 expert operators was available. Operator years of experience and data contribution in terms of whole scans and selected motion segments are presented in Table 1.

Table 1: Operator experience and data contribution

Operator	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Total
Experience (years)	0	1	14	10	2	6	3	5	8	15	10	7	-
Number of Scans	116	23	39	18	13	8	4	1	1	3	1	2	229
Number of Selected Motion Segments	214	35	67	26	20	12	7	3	2	3	2	5	396

2.2 Model Architecture

Our model needed to factor in that besides showing different proficiency levels during ultrasound scanning, operators also demonstrate personal scanning styles. Therefore, a naive network design might easily be biased to learn operator personal styles, in addition to the patterns that distinguish skill levels. One solution to this problem is to include a constraint in the network to reduce the influence of learned features from personal styles while preserving the influence of features from differentiating skill levels. To achieve this objective, we draw inspiration from recent success [6] in the domain adaptation field and introduce a domain classifier branch into our network.

As shown in Fig.1, our model consists of three parts; (1) a CNN feature extractor, (2) a skill level predictor, and (3) an operator classifier.

For motion feature extraction, we chose to implement a one-dimensional version of the 18-layer ResNet [8] for motion feature extraction. By introducing shortcut connections and residual blocks, ResNet [8] greatly alleviates gradient vanishing/exploding and has shown good performance on classification problems.

During the training phase, we expect the network to minimize the loss of skill level predictor while maximizing the loss of the operator classifier. Let $M = \{m_1, m_2, ..., m_N\}$ denote the probe motion dataset that has N motion segments. We use a neural network $f_v(m; \theta_v)$ with learnable parameters θ_v to

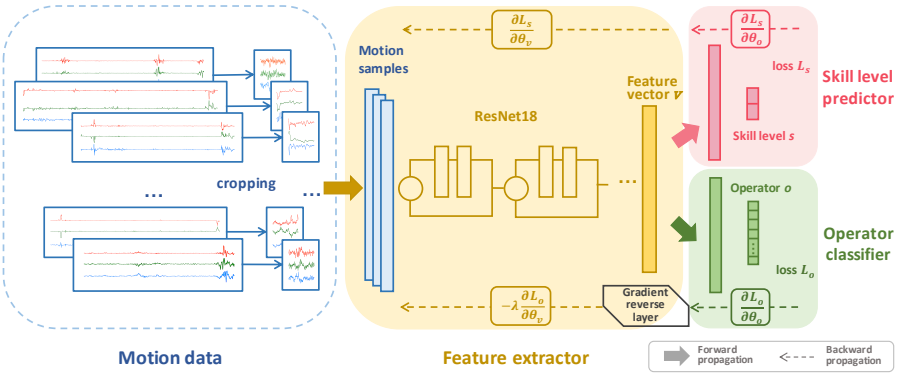


Fig. 1: The proposed framework for classifying operator skill levels. Motion segments are cropped into fixed sized samples before feeding into the feature extractor. During training, we minimize the loss of skill level predictor while at the same time maximize the loss of the operator classifier.

learn a mapping from each motion data m_i to a feature vector v_i. The feature vector is mapped to the skill level s by a skill level predictor $f_s(v;\theta_s)$ with learnable parameters θ_s. The skill level predictor is trained with loss function $L_s(f_s(f_v(m;\theta_v);\theta_s))$. For this we chose a binary cross-entropy loss. In the other branch, the operator classifier $f_o(v;\theta_o)$ tries to distinguish different operators $o_i \in O = \{o_1, o_2, ..., o_C\}$ from motion features where there are C operators. The parameters of the classifier are denoted as θ_o. The operator classifier is trained with loss function $L_o(f_o(f_v(m;\theta_v);\theta_o))$, which is a multinomial cross-entropy loss. Therefore, our model is trained with a joint loss function:

$$L(\theta_v, \theta_s, \theta_o) = L_s(f_s(f_v(m;\theta_v);\theta_s)) - \lambda L_o(f_o(f_v(m;\theta_v);\theta_o)), \tag{1}$$

where λ is the coefficient that balances the trade-off between two losses. During the training process, λ is gradually increased from 0 to 1 in order to limit the impact of operator loss at the early stage. We follow the strategy used in [6] to update λ:

$$\lambda = \frac{2}{1 + e^{-10p}} - 1, \tag{2}$$

where p represents the progress rate. Consider each training epoch $k \in \{1, 2, ..., K\}$, we feed a subset of motion data $M' = \{m_i, ..., m_j\}$ into the network. At the kth epoch, the progress p when feeding the motion data m_i is calculated by Eq. 3:

$$p = \frac{i + k \times (j - i + 1)}{K \times (j - i + 1)}. \tag{3}$$

2.3 Model Training

The duration of scanning varies from scan to scan. Therefore, for each epoch, we randomly cropped a fixed sized sample from each motion data segments and fed

the samples into our network. In the testing phase, the operator classifier was removed. For each motion segment, we sequentially cropped the data into small samples and predicted the skill level for each sample with the trained model. The final skill level prediction for each motion segment was given by a majority voting among the predictions of all samples within the segment (the mode of all labels).

The neural network was implemented with PyTorch framework and was trained on Nvidia GTX 1080Ti GPUs. The network was optimized with Adam Optimizer. The learning rate was set to 0.001 and we choose 32 as the batch size. We performed a parameter sweep over sample length and chose $2s$ as it was empirically the best performing time length.

2.4 Comparative Classification Methods

Most prior work in sonographer skill assessment using hand motion uses predefined features (path length, completion time, and points scanned, etc.) [16, 21]. Due to safety and privacy reasons in the pregnancy scan room, we could not use cameras or electromagnetic motion trackers as in some works. As a result, methods from prior work are not directly comparable to our data. Instead, we implemented an SVM which takes the statistical motion features as input to predict skill levels for comparison. We use the tsfresh [1] package to compute time series features and use the feature selection function to select the N most important features.

We also implemented an 18-layer 1D ResNet with skill level predictor only to show the effectiveness of our operator classifier branch.

3 Results and Discussion

We evaluated our proposed framework on the dataset mentioned in Section 2.1. Motion segments from the same subject could only be in one of the training, validation, or testing set. Detailed dataset partitioning is listed in Table 2.

Table 2: Dataset Description

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
Train	141	23	43	17	14	10	3	3	2	2	2	1
Validation	34	4	10	4	2	1	2	0	0	0	0	0
Test	39	8	14	5	4	1	2	0	0	1	0	4
Total	214	35	67	26	20	12	7	3	2	3	2	5

Operator skill classification results with comparison to a traditional SVM-based approach and an 18-layer 1D ResNet are presented in Table 3. We report model accuracy as well as precision, recall, and F1 score for expert group and

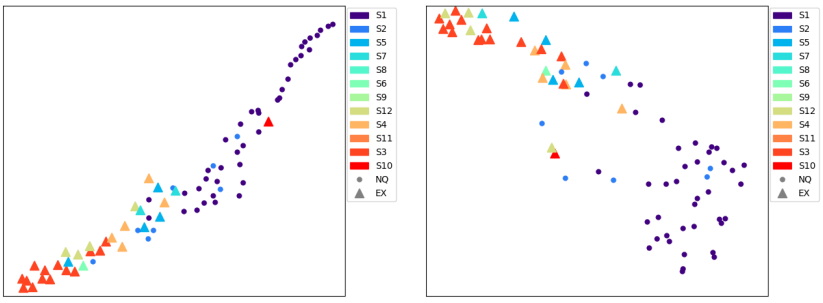
Table 3: Results of operator skill level classification with different models using motion data when scanning the brain. "EX" and "NQ" denotes expert and newly qualified respectively. "Macro" and "Weighted" refers to the mean and weighted mean for each evaluation metric.

		Precision	Recall	F1 Score	Accuracy
1D ResNet18	EX	0.88	0.71	0.79	0.85
	NQ	0.83	**0.94**	0.88	
	Macro	0.86	0.82	0.83	
	Weighted	0.85	0.85	0.84	
1D ResNet18 with operator loss	EX	**0.91**	**0.97**	**0.94**	**0.95**
	NQ	**0.98**	**0.94**	**0.96**	
	Macro	**0.94**	**0.95**	**0.95**	
	Weighted	**0.95**	**0.95**	**0.95**	
SVM	EX	0.67	0.65	0.66	0.75
	NQ	0.80	0.81	0.80	
	Macro	0.73	0.73	0.73	
	Weighted	0.75	0.75	0.75	

newly qualified group. Mean and weighted mean for each metrics are also calculated. Our proposed framework achieves a 95% accuracy for classifying skill level groups. From the results we observe that convolutional neural networks obtain better prediction scores than SVM. By adding the operator classifier branch for domain adaptation during training, the model is constrained to learn operator-invariant skill related features of skill level prediction. The improvement is noticeable in all evaluation metrics.

To further explore the difference of feature embeddings between the model with or without operator classifier, we visualize the features from the last convolutional layer using t-distributed stochastic neighbor embedding (t-SNE) with same parameters [10]. In Fig. 2, newly qualified level and expert level are presented with dot markers and triangle markers respectively. The years of operator experience is mapped to a different color where blue shows fewer years of experience and red shows more years of experience. As shown in the figure, both models successfully divided the motion data into two skill level groups. Figure 2a also shows the trend that the projected motion features of less experienced to more experienced operators distributed from the top right corner to the bottom left corner. In the meantime, the projected motion features from the same operator are more scattered in Fig. 2b than in Fig. 2a, which demonstrates the effectiveness of the operator classifier.

We also perform a set of leave-one-operator-out experiments to test the model performance on the unseen sonographer's motion data. For the SVM, we report the test result on the model with the best validation accuracy. For 1D ResNet and our proposed framework, we took five models after converging, and report mean as well as the standard deviation of the accuracy. As listed in Table 4, the



(a) 1D ResNet (b) 1D ResNet with operator classifier

Fig. 2: tSNE visualization of feature embedding of the last convolutional layer (best viewed in color).

Table 4: Leave One Operator Out Experiment Result

Leave one out operator	Accuracy (%) 1D ResNet18	1D ResNet18 with operator loss	SVM	Number of segments
S1	4.60 ± 1.46	**83.93** ± 16.11	25.70	214
S2	47.14 ± 9.86	**91.43** ± 6.52	34.29	35
S3	46.02 ± 6.06	**89.25** ± 8.09	41.79	67
S4	44.87 ± 9.84	46.92 ± 33.92	**50.00**	26
S5	**29.17** ± 6.07	26.00 ± 20.83	25.00	20
S6	56.94 ± 8.89	**66.67** ± 5.27	58.33	12
S7	7.14 ± 10.90	**82.85** ± 5.72	14.29	7
S8	38.89 ± 29.91	**86.67 ± 26.67**	66.67	3
S9	25.00 ± 25.00	**40.00 ± 48.99**	0.00	2
S10	0.00 ± 0.00	**73.34** ± 38.87	0.00	3
S11	**100.00 ± 0.00**	70.00 ± 40.00	50.00	2
S12	**86.67 ± 14.91**	60.00 ± 37.95	**80.00**	5

proposed model achieves the best mean accuracy on 5 out of 7 operators with more than 5 available motion segments. The accuracy improvement is significant. Without the operator loss, the models learn features with a large amount of operator variance. As a result, the models perform badly on the unseen operator. For the newly qualified (NQ) group (S1 and S2), the proposed model achieves 83.93% and 91.43% of mean accuracy when leaving out each operator, with comparison to the baseline model of 4.60% and 47.14%. Note that S1 contributes 86% of the data to the NQ group and 54% of the data to the whole dataset. Although taking out S1 would significantly reduce the training data and result in large data imbalance between two skill levels, the model still gives prediction with high accuracy. Now we look at the expert (EX) group with more than 5 motion segments (S3 to S7). S3 contributes 45.6% of the data to EX group, although taking out S3 makes the training set even more unbalanced between

the NQ and EX groups (from 62.9% vs. 37.1% to 75.7% vs. 24.3%), the model is still able to correctly predict an average of 89.25% of the S3 data as expert level. S4 is one experienced operator but classified as newly qualified more often, suggesting the scanning actions are similar to the latter category. This behavior, as an exception, needs to be further investigated in future work. For operator S5, the accuracy is low across all models, which may be caused by the fact that S5 only has 2 years of experience and the way that S5 scans is not yet approaching to the expert level. For the remaining operators (S8 to S12), their data contribution is very small therefore strong conclusions can not be drawn from the test results.

4 Conclusion and Future Work

We have presented a deep learning framework to model ultrasound probe motion for differentiating operator skill levels. The proposed model is evaluated on probe motion data that we collected from routine second-trimester fetal ultrasound scans undertaken by operators of different skill levels. The experiments show that the proposed model is capable of learning operator-invariant probe motion features that well distinguish operator skill levels and produce skill level prediction with high accuracy.

The proposed framework was designed as a model to support non-specialists or new trainees to assess their relative skill to an expert and could be used to establish when they reach expert competency in ultrasound scanning. It does not however identify areas of improvement which would be important for a trainee to progress. An interesting next step would be to see how to change the network to provide feedback on how to improve skills. The model might also be useful to assess the value of an assistive technology for a trainee, namely, when a trainee uses an assistive technology does their skill level approximate that of an expert. Our current work was limited by available data and data imbalance has been noted as an issue because we use data captured in a real-world setting. The model architecture is specifically designed to address data imbalance by forcing the model to learn operator-invariant features. However, it would be interesting to see how the model behaved on a larger balanced dataset.

Acknowledgement

We acknowledge the ERC (ERC-ADG-2015 694581, project PULSE), EPSRC (EP/M013774/1, Project Seebibyte), and the NIHR Oxford Biomedical Research Centre.

References

1. Tsfresh: Time series feature extraction based on scalable hypothesis tests. https://tsfresh.readthedocs.io/en/latest/

2. Ahmidi, N., Gao, Y., Béjar, B., Vedula, S.S., Khudanpur, S., Vidal, R., Hager, G.D.: String motif-based description of tool motion for detecting skill and gestures in robotic surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 26–33. Springer (2013)
3. Ahmidi, N., Ishii, M., Fichtinger, G., Gallia, G.L., Hager, G.D.: An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool-motion data. In: International forum of allergy & rhinology. vol. 2, pp. 507–515. Wiley Online Library (2012)
4. Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., Heng, P.A.: Standard plane localization in fetal ultrasound via domain transferred deep neural networks. IEEE journal of biomedical and health informatics **19**(5), 1627–1636 (2015)
5. Cox, B., Beard, P.: Imaging techniques: Super-resolution ultrasound. Nature **527**(7579), 451 (2015)
6. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495 (2014)
7. Hatala, R., Cook, D.A., Brydges, R., Hawkins, R.: Constructing a validity argument for the objective structured assessment of technical skills (osats): a systematic review of validity evidence. Advances in Health Sciences Education **20**(5), 1149–1175 (2015)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Kumar, R., Jog, A., Malpani, A., Vagvolgyi, B., Yuh, D., Nguyen, H., Hager, G., Chen, C.C.G.: Assessing system operation skills in robotic surgery trainees. The International Journal of Medical Robotics and Computer Assisted Surgery **8**(1), 118–124 (2012)
10. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
11. Salomon, L., Alfirevic, Z., Berghella, V., Bilardo, C., Hernandez-Andrade, E., Johnsen, S., Kalache, K., Leung, K.Y., Malinger, G., Munoz, H., et al.: Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound in Obstetrics & Gynecology **37**(1), 116–126 (2011)
12. Salomon, L., Alfirevic, Z., Bilardo, C., Chalouhi, G., Ghi, T., Kagan, K., Lau, T., Papageorghiou, A., Raine-Fenning, N., Stirnemann, J., et al.: Isuog practice guidelines: performance of first-trimester fetal ultrasound scan. Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in Obstetrics and Gynecology **41**(1), 102 (2013)
13. University of Oxford: PULSE: Perception ultrasound by learning sonographic experience. https://https://www.eng.ox.ac.uk/pulse/
14. Vedula, S.S., Ishii, M., Hager, G.D.: Objective assessment of surgical technical skill and competency in the operating room. Annual review of biomedical engineering **19**, 301–325 (2017)
15. Vrachnis, N., Papageorghiou, A.T., Bilardo, C.M., Abuhamad, A., Tabor, A., Cohen-Overbeek, T.E., Xilakis, E., Mates, F., Johnson, S.P., Hyett, J.: International society of ultrasound in obstetrics and gynecology (isuog)-the propagation of knowledge in ultrasound for the improvement of ob/gyn care worldwide: experience of basic ultrasound training in oman. BMC medical education **19**(1), 434 (2019)
16. Zago, M., Sforza, C., Mariani, D., Marconi, M., Biloslavo, A., La Greca, A., Kurihara, H., Casamassima, A., Bozzo, S., Caputo, F., et al.: Educational impact of

hand motion analysis in the evaluation of fast examination skills. European Journal of Trauma and Emergency Surgery pp. 1–8 (2019)
17. Zappella, L., Béjar, B., Hager, G., Vidal, R.: Surgical gesture classification from video and kinematic data. Medical image analysis **17**(7), 732–745 (2013)
18. Zia, A., Essa, I.: Automated surgical skill assessment in rmis training. International journal of computer assisted radiology and surgery **13**(5), 731–739 (2018)
19. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Clements, M.A., Essa, I.: Automated assessment of surgical skills using frequency analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 430–438. Springer (2015)
20. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Ploetz, T., Clements, M.A., Essa, I.: Automated video-based assessment of surgical skills for training and evaluation in medical schools. International journal of computer assisted radiology and surgery **11**(9), 1623–1636 (2016)
21. Ziesmann, M.T., Park, J., Unger, B., Kirkpatrick, A.W., Vergis, A., Pham, C., Kirschner, D., Logestty, S., Gillman, L.M.: Validation of hand motion analysis as an objective assessment tool for the focused assessment with sonography for trauma examination. Journal of Trauma and Acute Care Surgery **79**(4), 631–637 (2015)